Theor Appl Genet (2005) 110: 425–431
DOI 10.1007/s00122-004-1833-3

ORIGINAL PAPER

C. Lopez · B. Piégu · R. Cooke · M. Delseny
J. Tohme · V. Verdier

Using cDNA and genomic sequences as tools to develop SNP strategies in cassava (*Manihot esculenta* Crantz)

Received: 18 May 2004 / Accepted: 1 October 2004 / Published online: 14 January 2005

Abstract Single nucleotide polymorphisms (SNP) are the most abundant type of DNA polymorphism found in animal and plant genomes. They provide an important new source of molecular markers that are useful in genetic mapping, map-based positional cloning, quantitative trait locus mapping and the assessment of genetic distances between individuals. Very little is known on the frequency of SNPs in cassava. We have exploited the recently-developed collection of cassava expressed sequence tags (ESTs) to detect SNPs in the five cultivars of cassava used to generate the sequences. The frequency of intra-cultivar and inter-cultivar SNPs after analysis of 111 contigs was one polymorphism per 905 and one per 1,032 bp, respectively; totaling 1 each 509 bp. We have obtained further information on the frequency of SNPs in six cassava cultivars by analysis of 33 amplicons obtained from 3′ EST and BAC end sequences. Overall, about 11 kb of DNA sequence was obtained for each cultivar. A total of 186 SNPs (136 and 50 from ESTs and BAC ends, respectively) were identified. Among these, 146 were intra-cultivar polymorphisms, while 80 were inter-cultivar polymorphisms. Thus the total frequency of SNPs was one per 62 bp. This information will help to develop new strategies for EST mapping as well as their association with phenotypic characteristics.

Communicated by E. Guiderdoni

C. Lopez · B. Piégu · R. Cooke · M. Delseny · V. Verdier
Laboratoire Génome et Développement des Plantes,
UMR5096, CNRS-Université de Perpignan-Institut
de Recherche pour le Développement, 52 Av Paul Alduy,
66860 Perpignan Cedex, France

J. Tohme · V. Verdier (✉)
Biotechnology Research Unit Centro Internacional de Agricultura
Tropical, 6713 Cali, Colombia
E-mail: vverdier@univ-perp.fr
Tel.: +33-4-68661774
Fax: +33-4-68668499

Introduction

Genetic studies involving molecular genetic mapping, map-based positional cloning, quantitative trait locus mapping, estimates of genetic distances between individuals or the elucidation of the evolutionary history of populations require data from large sets of genetic markers. It has been customary to generate high density genetic maps based on the segregation data of RFLP, microsatellite or AFLP markers (Mohan et al. 1997; Phillips and Vasil 2001). Single nucleotide polymorphisms (SNPs) are a new class of markers that have recently attracted much interest. Given that SNPs can be identified in EST sequences, this type of polymorphism has the advantage of permitting the estimation of allele frequencies and association with interesting phenotypes. They are very stable markers compared with the tandem repeat markers. Recently several technologies that are amenable to automation have been developed for SNP discovery (Ching and Rafalski 2002; Gotoh and Oishi 2003; Pacey-Miller and Henry 2003; Schwarz et al. 2004).

Single nucleotide polymorphisms have been shown to be the most abundant type of molecular genetic markers in the genome (Cho et al. 1999) and are quickly becoming the marker of choice in agricultural research, especially for use in high-throughput marker-assisted breeding (Rafalski 2002). Based on different studies on animals and plants, it is necessary to sequence at least 200–500 bp of non-coding DNA on average to find a single non-coding SNP (ncSNP) and about 500–1,000 bp to locate a coding SNP (cSNP) (Brumfield et al. 2003). In humans it has been estimated that SNPs occur at a frequency of about one per 500–1,000 bp (Cooper et al. 1985; Wang et al. 1998). In plants, studies on the occurrence and nature of SNPs are beginning to receive considerable attention, particularly in *Arabidopsis*. In this plant more than 37,000 SNPs have been identified through the comparison of two accessions (Jander et al. 2002). In soybean, the presence of 280

SNPs in 143 amplicons totalling about 76.3 kb of DNA sequence has been reported (Zhu et al. 2003). Ching et al. (2002) reported the frequency of one ncSNP per 31 bp and 1 cSNP per 124 bp in 18 maize genes assayed in 36 inbred lines.

Expressed sequence tag databases are currently the fastest-growing and largest proportion of the available DNA sequence databases. The inherent redundancy in EST data makes them a potentially significant resource for the detection of SNPs. This resource has recently been used in a large-scale identification of SNPs in humans (Garg et al. 1999; Picoult-Newberg et al. 1999), *Arabidopsis* (Schmid et al. 2003), maize (Useche et al. 2001; Batley et al. 2003) and sugarcane (Grivet et al. 2003).

It has been shown that select amplicons in the non-coding regions, such as introns, 3′ untranslated regions (3′-UTRs) and BAC end sequences are a good source of data for SNP discovery and increase the frequency of detection of polymorphisms by up to threefold (Rafalski 2002; Zhu et al. 2003).

Cassava is a major calorie staple in the tropics and neotropics, providing a cheap source of dietary starch for over 700 million people in these regions of the world. In addition, starch produced from cassava roots and its industrially-produced derivatives are increasingly finding market niches in the food, baby formula, animal feed, paper, textile, alcohol and adhesive industries. Cassava is considered as an allopolyploid, with a high level of heterozygosity and suffers from inbreeding depression (Fregene et al. 1997). A molecular genetic linkage map of cassava has been constructed based principally on isoenzymes, RAPD and RFLP markers (Fregene et al. 1997) and recently updated with SSR markers (Mba et al. 2001). However, it remains difficult to obtain polymorphic markers in cassava. Only 40% of the available anonymous RFLP probes and 20% of the cDNA probes have shown polymorphism between the two parents of the cassava genetic map (Fregene et al. 1997; M. Fregene, personal communication). A number of other new resources have been generated over the last few years to improve the efficiency of cassava breeding. They include the detection of QTLs associated with agronomic characteristics (Okogbenin and Fregene 2003) and resistance to cassava bacterial blight (CBB) (Jorge et al. 2000, 2001), construction of BAC libraries (M. Fregene, personal communication), and the isolation of resistance gene candidates (RGCs) that can be used in marker-assisted selection as well as in map-based cloning of resistance genes (Lopez et al. 2003). Most recently we have generated a large collection of ESTs (Lopez et al. 2004). ESTs constitute an important tool for a better understanding of plant genome structure and gene expression and function. The development of an EST collection also provides an additional resource for the identification of new molecular markers and thus increases the density of gene markers on the genetic map. Here we have exploited cassava ESTs to detect SNPs in the cultivars used to generate the EST collection. We have obtained further information on the frequency of SNPs in cassava by analysis of 33 amplicons from 3′ EST and BAC end sequences in six cassava cultivars. This information will help to develop new strategies for the mapping of these ESTs and establish their association with phenotypic characteristics.

Materials and methods

For detection of SNPs, two approaches were followed: a bioinformatics-based analysis of the available ESTs, and laboratory experiments (PCR approach) on the non-coding sequences.

Identification of SNPs from the EST collection

A total of 11,954 ESTs were obtained from five cassava cultivars (Lopez et al. 2004). These were assembled into 1,875 tentative contigs (TC) and 3,175 singletons using the StackPack software (Miller et al. 1999). Polymorphisms were scored from the contigs obtained. Only contigs with four or more sequence reads (two different alleles, each represented by at least two independent sequence reads) were analyzed by inspection with polyBAYES (Marth et al. 1999). An SNP was considered as such only if each allele was confirmed by at least two reads in one cultivar.

The parameters for phrad were

"$PHRAP -new_ace -forcelevel 0 -ins_gap_ext −3 -del_gap_ext −3 -trim_score 20 -vecto_bound 0 -penalty −2 -gap_init −4 -gap_ext −3 -maxgap 30 -retain_duplicates $fseq 2 > /dev/null > $log "

For polyBAYES some parameters were left as the default values and others were specified as follows: anchorBaseQualityDefault 40-memberBaseQualityDefault 20-filterParalogs-paralogFilterMinimumBaseQuality 12-priorParalog 0.02-thresholdNative 0.75-screenSnps-considerAnchor-considerTemplateConsensus-prescreen Snps-preScreenSnpsMinimumBaseQuality 30-priorPoly 0.003-priorPoly2 0.99666-priorPoly3 0.00333-priorPoly4 $1e^{-05}$-priorPolyAC 0.1666-priorPolyAG 0.1666-priorPolyAT 0.1666-priorPolyCG 0.1666-priorPolyACG 0.25-priorPolyACT 0.25-thresholdSnp 0.5-maxTerms 50

Plant material and DNA extraction

Six different cultivars were used to determine the frequency of SNPs by a PCR-based approach. TMS30572 and CM2177-2 are the parents of the F_1 progeny used to construct the framework molecular genetic map of cassava (Fregene et al. 1997) while cultivars MCol1522 and Ecu72 are the parents of the segregating population used for the construction of a new genetic map anchored with microsatellite markers (CIAT 2002). The two other cultivars, MBra685 and SG107-35 are highly resistant to

cassava bacterial blight and were used to generate some of the EST collection.

Leaves from 1-month-old plants grown in greenhouses were collected in liquid nitrogen at CIAT (Cali, Colombia). DNA was extracted using the method described by Dellaporta et al. (1983).

Detection of SNPs by a PCR-based approach

Candidate ESTs tagging genes involved principally in disease resistance (resistance gene candidates, receptor-like proteins) and starch biosynthesis were selected from the collection (Table 1). These ESTs were generated from the 5′-ends of their genes. To obtain 3′-UTR sequences, clones were sequenced using the T7 primer and ABI Prism BigDye terminator sequencing chemistry (Applied Biosystems, Foster City, Calif., USA). Cycle sequencing reactions were carried out on plasmid DNA extracted as described (Lopez et al. 2004). Sequences were run on an ABI PRISM 3100 Genetic Analyzer (Applied Biosystems).

Based on the 3′ sequence information, gene-specific primer pairs were designed using the Primer3 program (http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi) as close to the 3′-end of the transcript as possible. The expected product size was 250–500 bp on average. BAC library screening with 12 RGC classes as probes allowed the identification of 42 BAC clones that were assembled into ten contigs and 19 singletons (Lopez et al. 2003). Most of the BAC ends were sequenced. Fifteen pairs of primers were designed from the 67 sequences obtained as described above with an expected product size of 300–500 bp (Table 2).

Testing PCR primers

All the primers were first used to amplify genomic DNA from one cultivar (notably TMS30572). Amplification reactions were performed in a 50 µl volume, containing 100 ng of DNA, 0.2 µM of each primer, 200 µM of each dNTP, 1× PCR buffer and 1.25 U *Taq* polymerase (AmpliTaq, Promega, Madison, Wisc., USA). PCR cycling conditions using an MJ Research thermocycler were 35 cycles with the following profile: 95°C for 45 s, 52°C for 45 s and 72°C for 1 min. Amplicons were visualized on a 1.5% agarose gel stained with ethidium

Table 1 Description of the ESTs used in the detection of SNPs in cassava. Similarities are reported based on the BLASTX analyses obtained from the 5′ sequence of the EST. The 3′ sequence was used to generate the primers used to detect the SNPs

Genbank accession	Similarity	Expected	Size (bp)	No of SNPs
CK642884	Methionine adenosyltransferase (tomato)	$3e^{-71}$	357	2
CK642931	Transketolase TKT1 precursor, chloroplast (pepper)	$1e^{-64}$	253	2
CK645247	Receptor-like protein kinase (rice)	$2e^{-78}$	339	2
CK645139	Hypothetical protein F24B22.150 (*Arabidopsis thaliana*)	$4e^{-71}$	194	1
CK644372	Probable disease resistance protein (*A. thaliana*)	$1e^{-35}$	366	
CK644249	5-Methyltetrahydropteroyltriglutamate-homocysteine S-methyltransferase (Madagascar periwinkle)	$7e^{-61}$	273	0
CK642228	1,4-Alpha-glucan branching enzyme (cassava)	$5e^{-08}$	324	6
CK642341	1,4-Alpha-glucan branching enzyme precursor, amyloplast (potato)	$5e^{-77}$	318	3
CK650910	Probable nematode-resistance protein (*A. thaliana*)	$2e^{-46}$	215	3
CK651476	Disease resistance protein RPS2 homolog T12H20.8 (*A. thaliana*)	$3e^{-19}$	312	5
CK644436	Chitinase class I (upland cotton)	$3e^{-36}$	259	5
CK644582	Probable WRKY-type DNA binding protein (*A. thaliana*)	$1e^{-26}$	137	10
CK644648	Resistance protein RGC2K (garden lettuce)	$4e^{-16}$	276	10
CK644959	No hit found	–	214	3
CK645381	NBS–LRR type resistance protein (rice)	$8e^{-36}$	364	24
CK645466	Chitinase precursor, basic (*A. thaliana*)	$1e^{-12}$	264	3
CK645404	Disease resistance protein EDS1 (*A. thaliana*)	$1e^{-47}$	159	6
CK644194	Resistance protein homolog RGC2a (garden lettuce)	$1e^{-08}$	346	2
CK648562	Disease resistence protein homolog F24J7.60 (*A. thaliana*)	0.001	315	3
CK642817	Disease resistance protein Cf-2.1 (currant tomato)	$4e^{-21}$	282	0
CK644727	Probable receptor-like protein kinase (*A. thaliana*)	$1e^{-26}$	177	1
CK645969	Resistance protein RGC2K (garden lettuce)	0.9	125	4
CK646080	Hypothetical protein (*A. thaliana*)	$6e^{-24}$	234	1
CK646023	No hit found	–	253	4
CK648035	TMV resistance protein N homolog F23E13.40 (*A. thaliana*)	$1e^{-08}$	220	0
CK649815	Probable nematode-resistance protein (*A. thaliana*)	$3e^{-40}$	339	1
CK644703	Disease resistance-like protein (*A. thaliana*)	0.00002	431	13
CK644318	Probable receptor-like protein kinase (*A.thaliana*)	$5e^{-93}$	290	5
CK643078	Disease resistance protein Cf-2.1 (currant tomato)	$1e^{-32}$	319	5
CK901350	H + -transporting ATPase, plasma membrane (*Prunus persica*)	$2e^{-19}$	156	5
CK901144	No hit found	–	212	2
RGC 7	Probable disease resistance protein RPS2 (*A. thaliana*)	$6e^{-41}$	534	3
RGC 10	NBS–LRR type resistance protein (rice)	$2e^{-53}$	290	2

Table 2 Description of the BAC end sequences used for the detection of SNPs in cassava

Code BAC	Similarity	Expect	size	No of SNPs
08H19	No hit found	–	216	12
26K13	No hit found	–	126	9
39P22	NBS–LRR type resistance protein (rice)	$5.00e^{-18}$	448	7
70J21	No hit found	–	185	2
75D2	No hit found	–	304	9
82N9	No hit found	–	230	4
84G19	Fim protein (garden snapdragon)	$2.00e^{-05}$	382	4
90N6	No hit found	–	400	0
121I15	No hit found	–	179	3

bromide. Products were directly sequenced using one of the PCR primers with BigDye terminator cycle sequencing (Perkin-Elmer Applied Biosystems) and analyzed on a Perkin-Elmer ABI Prism 3100 Genetic Analyzer. Sequences were edited and analyzed using Sequencher (Gene Codes, Ann Arbor, Mich., USA). Primers producing amplicons suitable for sequencing were used to analyze the other cassava cultivars.

SNP discovery

Sequences obtained from all cultivars were aligned using Sequencher (Gene Codes) and inspected visually for the detection of polymorphisms. Putative SNPs were accepted only if the sequence trace was of high quality.

Results

Identification of SNPs in the EST collection

Our first approach was global, using contigs constructed from our EST collection. Despite their reputation as low-quality sequences, ESTs represent a rich source of molecular information, as long as the quality values for individual bases are available. We have explored the use of cassava EST sequences (Lopez et al. 2004) for the identification of cSNPs using data from the 1,875 contigs obtained after assembly using the StackPack software (Miller et al. 1999). Among these, 964 contained four or more sequence reads. A conservative approach was followed so that only polymorphisms represented by two or more sequences were considered. The sequences were inspected for the presence of polymorphisms using polyBAYES software (Marth et al. 1999). This software uses trace quality values attributed by the base caller phred (Ewing et al. 1998) and thus takes into account the variability in sequence quality inherent in the EST approach. Among the contigs analyzed, 111 contained sequence variants (SNP plus indels) which could be divided into two types: those present within the same cultivar,—intra-cultivar SNPs—and those that discriminated between two or more cultivars—-inter-cultivar SNPs. We found 81 SNPs and 15 indels in the first category and 76 SNPs and five indels in the second (Table 3). The number of SNPs per contig varied from one to seven. The estimated frequencies were of one per 905 bp for the intra-cultivar SNPs and one per 1,032 bp for the inter-cultivar SNPs. Transitions (C/T or G/A and vice versa) were most common in both intra- and inter-cultivars (64 and 65% respectively, Table 3) than transversions (A/C, A/T, G/C or G/T and vice versa). In total the number of transitions was significantly higher than transversions (93 vs. 51, $P = 0.0005$). A greater number of indels were detected within cultivars (15) than between them (5, Table 3). Overall, 144 SNPs were detected, totalling 73,332 bp, thus giving a total of one SNP every 509 bp.

3′ sequencing of selected ESTs

The second method exploiting the EST sequences was a targeted approach, used after identifying candidate genes. The 3′ end sequences of these ESTs began with the polyA$^+$ tail, indicating that they include at least part of the 3′-UTR of each gene. For many of these 3′ sequences, similarity searches using BLASTX showed no significant similarity (data not shown), as was expected for untranslated or highly divergent sequences.

SNP detection derived from 3′ ESTs

Among the 81 primer pairs based on the 3′ sequences, 59 produced single bands (73%), 17 (21%) yielded no PCR products and 5 (6%) gave products with either multiple or weak bands. The 59 PCR products with single bands provided 33 good quality sequences (Table 4). Sequences of these amplicons obtained from six different

Table 3 Summary of cassava sequence variant analyses based on the EST collection

Type of SNP	Number of SNPs		
	Intra-cultivar	Inter-cultivar	Total
Transitions (%)	52 (64)	46 (65)	93 (65)
Transversion (%)	29 (36)	25 (35)	51 (35)
Indel	15	5	20
Total	81 (59)	71 (46)	165
No. of contigs analyzed	111		
Bases	73,320		

Table 4 Summary of cassava sequence variants analyses based on PCR assays

	Source of sequence			
	EST		BAC end	
No. of primer pairs designed	81		17	
No. of good quality sequences	33		9	
Bases	9,042		2,470	
Type of SNP	Intra-cultivar	Inter-cultivar	Intra-cultivar	Inter-cultivar
Transitions (%)	58 (56)	35 (54)	19 (44)	10 (67)
Transversion (%)	45 (44)	30 (46)	24 (56)	5 (33)
Indel	–	6	–	1
Total	103	65	43	15
	136		50	

cassava cultivars were aligned and SNPs detected. We found 103 SNPs within individual cultivars which were classified according to the nucleotide substitution as either transitions or transversions. There were 58 transitions (56%) and 45 transversions (44%). Sixty-five SNPs discriminated between at least two cultivars. For this type of SNP there were 35 transitions (54%) and 30 transversions (46%).

We detected six indels between cultivars. Overall, approximately 9 kb of sequence was obtained for each cultivar. A total of 136 SNPs were identified. Thus the frequency of SNPs was one per 66 bp (Table 4). The ESTs that showed the highest level of SNPs were CK644582 and CK644648 with 10 SNPs each, CK644703 (13 SNPs), and CK645381 (24 SNPs). ESTs CK644249, CK642817 and CK648035 had no SNPs. ESTs CK644727, CK646080, CK649815 and CK645139 presented only one SNP each (Table 1).

SNP detection derived from BAC end sequences

Seven primer pairs out of a total of 17 designed from BAC end sequences produced several bands while ten produced only one fragment. High quality sequence data were obtained from nine of the ten amplicons produced from these primers (Table 4). We found 43 intra-cultivar (86%) and 15 inter-cultivar SNPs (30%). Transitions represented 44 and 67% of the substitutions for the intra-cultivar and inter-cultivar SNPs, respectively. A higher percentage of transversions were found within (56%) than between the cultivars (33%). Overall, 50 SNPs were detected, totaling about 2.4 kb of sequence for each cultivar. The overall frequency of SNPs was one per 49 bases. The BAC end sequences that showed most polymorphism were 08H19 with 12 SNPs and sequences 26K13 and 75D2 with nine SNPs each (Table 2). The BAC end sequence from 90N6 shows no SNPs except for one indel. BAC end sequence 70J21 shows only two SNPs (Table 2).

Overall, for the 3′ EST and BAC end sequences we detected 186 SNPs from a total of 11.5 kb of sequence for each cultivar. The mean occurrence of SNPs was one per 62 bp. The frequency of transversions (46%) is almost equal to that of transitions (54%, $P = 0.3$).

Discussion

We have exploited the recent EST collection generated from cassava (Lopez et al. 2004) to explore the frequency of SNPs in this crop. The Stackpack (Miller et al. 1999) criteria used to construct the EST contigs led to tags for allelic sequences being incorporated into one contig. This information was used to study genetic and allelic diversity in cassava and, more specifically, to detect SNPs within and between cassava cultivars. Taking into account the poor quality data inherent in single-pass EST sequencing, SNPs were only retained if at least two reads were available for each SNP. These polymorphisms can be considered as candidate SNPs. While it will be necessary to validate these by direct sequencing, similar studies carried out in EST collections in maize and barley have shown that the majority of predicted SNPs represent true genetic variation (Batley et al. 2003; Kota et al. 2003). The frequency of cSNPs in cassava was one polymorphism per 509 bp.

Considering the high level of heterozygosity in cassava, we expected to find a greater number of contigs with two alleles for each cultivar. The actual numbers are relatively low, probably because 5′-EST sequencing gives a very high proportion of coding regions where polymorphism is rare. Higher levels of SNPs should be found outside the coding regions. We used two approaches to test this possibility: sequencing of the 3′-ends of selected ESTs to obtain 3′ non-coding regions and amplification of fragments based on BAC end sequences, which can be considered to be more or less random genomic sequences. These two methods allowed the evaluation of the presence of ncSNPs in several cultivars.

Most of the primers designed from the 3′ ESTs did not yield high-quality sequence data. Inspection of these PCR products on agarose gels showed that the amplification was successful and apparently specific (only a single PCR product was visualized). Sequence analyses showed the presence of a heterogeneous population of sequences of the same size, indicating the presence of several members of the same multigene family.

Our results show a frequency of ncSNPs approximately seven times greater than from the global

EST analysis, indicating the importance of non-coding regions in the detection of a greater number of polymorphisms. It has been reported that the non-coding regions can increase the frequency of polymorphism by up to threefold (Rafalski 2002).

We detected indels in the EST collection. Direct sequencing of genomic PCR products did not result in the detection of intra-cultivar indels in heterozygous individuals, as observed in the 3′ UTR or BAC end sequences.

The value of one SNP per 62 bp is close to those obtained for other crops. For example, Ching et al. (2002) reported the presence of one polymorphism per 31 bp in non-coding regions and one per 124 bp in coding regions based on the analysis of 18 maize genes in 36 inbred lines. In soybean, a total of 280 SNPs were identified in 143 DNA fragments, representing an approximate frequency of one SNP per 270 bp (Zhu et al. 2003). Analyses of polymorphism in the *Adh* genes from a sugarcane EST database revealed a mean occurrence of one SNP every 122 bp (Grivet et al. 2003). The values are similar between different species, although the frequency of SNPs in each species is slightly different. This may simply be due to the kind of sequence data used to generate the SNPs, and to the reproduction mode of the species. In this regard, the frequency of SNP in cassava genome is comparable to that of corn, another outcrossing species.

It is obvious that non-coding regions can accumulate a greater number of polymorphisms and that not all genes accumulate SNPs at the same rate. We observed two groups of genes: those containing a relatively high number of SNPs (more than 6) and another with few or no SNPs. There was no correlation between the number of SNPs and the size of the region sequenced, some of the shorter EST contigs giving greater numbers of SNPs. Thus, there are genes or regions of genes that evolve more rapidly. Constraints are obviously different between coding and non-coding regions and, even within coding sequences, there is a strong bias towards synonymous mutations in regions that are important for the function of the protein. It is possible that this is the case for ESTs CK644249, CK642817 and CK648035. On the other hand, we determined that the sequenced region of EST CK644703 corresponds to the LRR region present in the R-genes. This is known to be a highly variable sequence since it is involved in the direct or indirect interaction with the Avr protein of the pathogen (Ellis et al. 1999). This may explain the relatively high number of SNPs present in this gene.

Single nucleotide polymorphisms may differentiate between allelic polymorphism within a single variety (intra-cultivar SNPs) or between cultivars (inter-cultivar SNPs) (Batley et al. 2003). The SNPs we detected that differentiate cultivars can be used directly for mapping. We also detected intra-cultivar SNPs. It will be interesting to test if all allelic polymorphisms within a single variety are expressed in each cultivar. The identification of SNPs in barley using ESTs has led to the detection of 3,069 candidate inter-varietal SNPs and 3,377 hypothetical intra-varietal SNPs (Kota et al. 2003). The presence of an intra-varietal polymorphism was explained by the presence of a background of heterozygosity within the barley "inbred" lines or by the possibility that the cultivars used for cDNA library construction may consist of mixtures of discernable genotypes.

Cassava is considered to be an allopolyploid although its diploid ancestors are not known (Fregene et al. 1997). It presents a high level of heterozygosity and suffers from inbreeding depression, making it difficult to obtain homozygous lines. These characteristics of cassava may explain the high rate of intra-cultivar SNPs obtained in this study. However, the ancient polyploid origin of the crop makes the analyses of these polymorphisms complicated; one should indeed expect that at least some of the SNPs would be found in paralogous loci within the genome, rather than being heterozygous at a given locus. Nonetheless, we did not detect more than two alleles within a particular cultivar, suggesting that several paralogs are not present in the same contig. Only mapping of these SNPs could determinate clearly whether the polymorphism is between paralogs or alleles.

Cassava is considered to be a species with a recent evolutionary origin (Roger and Appan 1973). Finding DNA sequence variation has been difficult (Olsen et al. 1998) and has made studies of diversity and phylogeography rather difficult. This study provides new tools for genetic mapping in cassava. The SNPs were detected in cultivars used as the parents of cassava genetic mapping populations. Thus, this information can be used directly for mapping the ESTs using more cost-effective, high-throughput SNP assays in an automated fashion, such as DHPLC. The genes detected in our study showing higher numbers of SNPs represent new markers useful for genetic diversity studies, phylogeography, germplasm characterization, mapping or other studies based on molecular markers.

Acknowledgements We are grateful to Chike Mba and Olivier Panaud for critically reading the manuscript. Thanks to Christel Llauro-Berger and Michèle Laudié for sequencing. All the sequencing was performed by the facilities of the Montpellier Languedoc Rousillon Genopole. Camilo Lopez was supported by a doctoral fellowship awarded by the IRD.

References

Batley J, Barker G, O'Sullivan H, Edwards KJ, Edwards D (2003) Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. Plant Physiol 132:84–91

Brumfield RT, Beerli P, Nickerson DA, Edwards SV (2003) The utility of single nucleotide polymorphisms in inferences of population history. Trends Ecol Evol 18:249–256

Ching A, Rafalski A (2002) Rapid genetic mapping of ESTs using SNP pyrosequencing and indel analysis. Cell Mol Biol Lett 7:803–810

Ching A, Caldwell KS, Jung M, Dolan M, Smith OS, Tingey S, Morgante M, Rafalski AJ (2002) SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genet 3:19

Cho RJ, Mindrinos M, Richards DR, Sapolsky RJ, Anderson M, Drenkard E, Dewdney J, Reuber TL, Stammers M, Federspiel N, Theologis A, Yang WH, Hubbell E, Au M, Chung EY, Lashkari D, Lemieux B, Dean C, Lipshutz RJ, Ausubel FM, Davis RW, Oefner PJ (1999) Genome-wide mapping with biallelic markers in *Arabidopsis thaliana*. Nat Genet 23: 203–207

CIAT (2002) Annual Report, Centro Internacional de Agricultura Tropical, Cali, Colombia

Cooper DN, Smith BA, Cooke HJ, Niemann S, Schmidtke J (1985) An estimate of unique DNA sequence heterozygosity in the human genome. Hum Genet 69:201–205

Dellaporta SL, Wood J, Hicks JB (1983) A plant DNA mini-preparation: version II. Plant Mol Biol Rep 1:19–21

Ellis JG, Lawrence GJ, Luck JE, Dodds PN (1999) Identification of regions in alleles of the flax rust resistance gene *L* that determine differences in gene-for-gene specificity. Plant Cell 11:495–506

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred I accuracy assessment. Genome Res 8:175–185

Fregene M, Angel F, Gomez R, Rodriguez F, Chavarriaga P, Roca W, Tohme J, Bonierbale M (1997) A molecular genetic map of cassava (*Manihot esculenta* Crantz). Theor Appl Genet 95:431–441

Garg K, Green P, Nickerson DA (1999) Identification of candidate coding region single nucleotide polymorphisms in 165 human genes using assembled expressed sequence tags. Genome Res 9:1087–1092

Gotoh K, Oishi M (2003) Screening of gene-associated polymorphisms by use of in-gel competitive reassociation and EST (cDNA) array hybridization. Genome Res 13:492–495

Grivet L, Glaszmann JC, Vincentz M, da Silva F, Arruda P (2003) ESTs as a source for sequence polymorphism discovery in sugarcane: example of the *Adh* genes. Theor Appl Genet 106:190–197

Jander G, Norris SR, Rounsley SD, Bush DF, Levin IM, Last RL (2002) *Arabidopsis* map-based cloning in the post-genome era. Plant Physiol 129:440–450

Jorge V, Fregene MA, Duque MC, Bonierbale MW, Tohme J, Verdier V (2000) Genetic mapping of resistance to bacterial blight disease in cassava (*Manihot esculenta* Crantz). Theor Appl Genet 101:865–872

Jorge V, Fregene M, Velez CM, Duque MC, Tohme J, Verdier V (2001) QTL analysis of field resistance to *Xanthomonas axonopodis* pv . *manihotis* in cassava. Theor Appl Genet 102:564–571

Kota R, Rudd S, Facius A, Kolesov G, Thiel T, Zhang H, Stein N, Mayer K, Graner A (2003) Snipping polymorphisms from large EST collections in barley (*Hordeum vulgare* L.). Mol Genet Genomics 270:24–33

Lopez CE, Zuluaga AP, Cooke R, Delseny M, Tohme J, Verdier V (2003) Isolation of resistance gene candidates (RGCs) and characterization of an RGC cluster in cassava. Mol Genet Genomics 269:658–671

Lopez CE, Jorge V, Piegu B, Mba Ch, Cortes D, Restrepo S, Soto M, Laudié M, Berger Ch, Cooke R, Delseny M, Tohme J, Verdier V (2004) A unigene catalogue of 5,700 expressed genes in cassava. Plant Mol Biol (in press)

Marth GT, Korf I, Yandell MD, Yeh RT, Gu Z, Zakeri H, Stitziel NO, Hillier L, Kwok PY, Gish WR (1999) A general approach to single-nucleotide polymorphism discovery. Nat Genet 23:452–456

Mba REC, Stephenson P, Edwards K, Melzer S, Nkumbira J, Gullberg U, Apel K, Gale M, Tohme J, Fregene M (2001) Simple sequence repeat (SSR) markers survey of the cassava (*Manihot esculenta* Crantz) genome: towards an SSR-based molecular genetic map of cassava. Theor Appl Genet 102:21–31

Miller RT, Christoffels AG, Gopalakrishnan C, Burke J, Ptitsyn AA, Broveak TR, Hide WA (1999) A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. Genome Res 9:1143–1155

Mohan M, Nair S, Bhagwat A, Krishna TG, Yano M, Bhatia CR, Sasaki T (1997) Genome mapping, molecular markers and marker-assisted selection in crop plants. Mol Breed 3:87–103

Okogbenin E, Fregene M (2003) Genetic mapping of QTLs affecting productivity and plant architecture in a full-sib cross from non-inbred parents in Cassava (*Manihot esculenta* Crantz). Theor Appl Genet 107:1452–1462

Olsen KM, Hernandez M, Schaal BA (1998) A survey of DNA sequence variation in cassava and other *Manihot* species. In: Cassava biotechnology: proceedings of the IVth international scientific meeting of the cassava biotechnology network. EMBRAPA-Recursos Geneticos e Biotecnologia, Brasilia

Pacey-Miller T, Henry R (2003) Single-nucleotide polymorphism detection in plants using a single-stranded pyrosequencing protocol with a universal biotinylated primer. Anal Biochem 317:166–170

Phillips RL, Vasil IK (2001) DNA-based markers in plants. Kluwer, Dordrecht

Picoult-Newberg L, Ideker TE, Pohl MG, Taylor SL, Donaldson MA, Nickerson DA, Boyce-Jacino M (1999) Mining SNPs from EST databases. Genome Res 9:167–174

Rafalski A (2002) Applications of single nucleotide polymorphisms in crop genetics. Curr Opin Plant Biol 5:94–100

Roger DJ, Appan SG (1973) *Manihot* and *Manihotoides* (Euphorbiaceae): a computer assisted study. Monograph no. 13, flora neotropica. Hafner, New York

Schmid KJ, Sorensen TR, Stracke R, Torjek O, Altmann T, Mitchell-Olds T, Weisshaar B (2003) Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in *Arabidopsis thaliana*. Genome Res 13:1250–1257

Schwarz G, Baumler S, Block A, Felsenstein FG, Wenzel G (2004) Determination of detection and quantification limits for SNP allele frequency estimation in DNA pools using real time PCR. Nucleic Acids Res 32:e24

Useche FJ, Gao G, Harafey M, Rafalski A (2001) High-throughput identification, database storage and analysis of SNPs in EST sequences. Genome Inform Ser Workshop Genome Inform 12:194–203

Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topaloglou T, Hubbell E, Robinson E, Mittmann M, Morris MS, Shen N, Kilburn D, Rioux J, Nusbaum C, Rozen S, Hudson TJ, Lander ES et al (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 280:1077–1082

Zhu YL, Song QJ, Hyten DL, Van Tassell CP, Matukumalli LK, Grimm DR, Hyatt SM, Fickus EW, Young ND, Cregan PB (2003) Single-nucleotide polymorphisms in soybean. Genetics 163:1123–1134